Efficient Low Complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G

Karimidehkordi, Ali; Pedersen, Klaus Ingemann; Mahmood, Nurul Huda; Gerardino, Guillermo Andrés Pocovi; E. Mogensen, Preben

Published in:
2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)

DOI (link to publication from Publisher):
10.1109/VTCSpring.2019.8746407

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Efficient Low Complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G

Ali Karimi[1], Klaus I. Pedersen[1,2], Nurul Huda Mahmood[1], Guillermo Pocovi [2], and Preben Mogensen[1,2]
[1]*Wireless Communications Networks (WCN) Section, Department of Electronic Systems, Aalborg University, Denmark.*
[2]*Nokia-Bell Labs, Aalborg, Denmark.*
alk@es.aau.dk

***Abstract*—We address the problem of resource allocation and packet scheduling for a mixture of ultra-reliable low-latency communication (URLLC) and enhanced mobile broadband (eMBB) traffic in a fifth generation New Radio (5G NR) networks. A novel resource allocation method is presented that is latency, control channel, hybrid automatic repeat request (HARQ), and radio channel aware in determining the transmission resources for different users. This is of high importance for the scheduling of URLLC users in order to minimize their latency, avoid unnecessary costly segmentation of URLLC payloads over multiple transmissions, and benefit from radio channel aware multi-user diversity mechanisms. The performance of the proposed algorithm is evaluated with an advanced 5G NR compliant system level simulator with a high degree of realism. Simulation results show promising gains of up to 98% latency improvement for URLLC traffic and 12% eMBB end-user throughput enhancement as compared to conventional proportional fair scheduling.**

***Index Terms*—5G NR, URLLC, Packet Scheduling**

I. Introduction

The fifth generation New Radio (5G NR) is set to support different services such as ultra-reliable low-latency communications (URLLC) and enhanced mobile broad (eMBB) [1]. For URLLC, various classes with different quality of service (QoS) requirements are defined by 3GPP, where one of the most stringent service target is one millisecond (msec) latency at 99.999% reliability [2]. An overview of communication theoretic principles of URLLC can be found in [3], [4]. A flexible multi-service capable frame structure has been studied in [5]. Several contributions in the literature have also studied various resource allocation techniques to enhance the performance of URLLC in 5G NR. The authors in [6] study the problem of user (UE) selection and scheduling for URLLC, where only one UE is scheduled in each transmission time interval. In [7], [8], the authors formulated a multi-dimensional 0-1 Knapsack problem for low-latency communications to select and drop delayed packets from the network. It has been shown in [9] that wide-band allocation maximizes the outage capacity of URLLC and dynamic multiplexing of URLLC and eMBB significantly improves the spectral efficiency. Dynamic link adaptation and multiplexing of URLLC and eMBB traffic on a shared channel were studied in [10], [11]. Finally, several pre-emptive scheduling schemes for multiplexing of URLLC and eMBB traffic are proposed in [12], [13].

In this paper, we present additional scheduler advancements as compared to earlier published studies. For scheduling of the high-priority UEs, we propose a resource allocation scheme that is payload and control channel aware, and exploits the radio channel time-frequency variations. The payload awareness is incorporated in the scheduler by favouring scheduling of full URLLC payloads without segmenting those over multiple transmissions. At most one UE per URLLC scheduling interval is subjected to segmentation, limited to the UE with the minimum segmentation cost. Moreover, the buffering time of individual payloads are explicitly taken into account in the scheduling decisions, as compared to the latency target. The overhead from the physical layer control channel to signal the scheduling grant to the UEs is also explicitly incorporated in the presented resource allocation framework. Finally, the proposed scheduler also has an element of radio channel awareness to gain from multi-user diversity.

State-of-the-art 5G NR compliant multi-cell dynamic system level results are presented to demonstrate how the proposed solution performs under different load regimes. The results confirm that the proposed resource allocation algorithm improves the latency performance of URLLC users, and also enhances the end-user throughput for the eMBB users.

The rest of the paper is organized as follows: the system model and problem formulation are elaborated in Section II. Section III discusses the proposed packet scheduling algorithm. Simulation methodology and performance results are presented in Section IV. Finally, the study is concluded in Section V.

II. Setting the Scene

A. Basic System Model

We adopt the 5G NR specifications as outlined in [1], [14], focusing primarily on the downlink (DL) performance for frequency division duplexing (FDD) mode. The network consists of C cells forming a

three-sectorized deployment with 500 meters inter-site distance corresponding to the 3GPP urban macro (UMa) deployment [14]. A set of U URLLC and M eMBB UEs are randomly distributed over the entire network area. For each URLLC UE, bursts of small payloads of B bytes arrive at the network according to a Poisson point process with arrival rate of λ [payload/sec]. This traffic model is known as FTP3 in 3GPP [15].

Full buffer traffic with infinite payload size is assumed for eMBB UEs. In the t-th transmission time interval (TTI), the sets of active (with data) URLLC and eMBB UEs connected to cell c are denoted by $\mathbf{U}^{c,t}$ and $\mathbf{M}^{c,t}$, respectively.

Both eMBB and URLLC traffic are dynamically multiplexed on a shared channel, using orthogonal frequency division multiple access (OFDMA) with 30 kHz sub-carrier spacing. A short TTI size of 4 OFDM symbols (0.143 msec) and a physical resource block (PRB) resolution of 12 sub-carriers is assumed as the minimum time and frequency scheduling unit.

The base stations and users are each equipped with two transmit/receive antennas. UEs exploit linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver to suppress noise and received interference. Each UE periodically measures the channel and interference for each resource element (RE) and reports a frequency-selective channel quality indicator (CQI) per sub-channel of eight PRBs. The reported CQIs are subjected to processing delay before being applied at the network for DL transmission.

User-centric control channel transmission is assumed to indicate the scheduling grant of scheduled UEs [16]. Thus, whenever a user is scheduled, both a user-specific scheduling grant on the physical downlink control channel (PDCCH) and the actual transport block (data) on the physical downlink shared channel (PDSCH) are transmitted. The PDCCH size is dynamically adjusted based on the reported wide-band signal to interference plus noise ratio (SINR) value to guarantee low probability of failure. In line with [10], [16], the PDCCH is transmitted with aggregation level 1,2,4, or 8 depending on the experienced SINR at the UE, where the aggregation consumes 36 REs.

Dynamic link adaptation is applied for transmission of the PDSCH. As the CQI is subjected to reporting delay and other imperfections, the well-known outer loop link adaptation (OLLA) is applied to control the block error rate (BLER). In line with [10], [17], the OLLA offset is adjusted to achieve 1% and 10% BLER of the first PDSCH transmission for URLLC and eMBB, respectively. In case of packet failure, the UE will feed back a negative acknowledgement (NACK), and the corresponding hybrid automatic repeat request (HARQ) retransmission is scheduled by the network. Asynchronous HARQ retransmission with Chase combining and a maximum of six retransmissions are assumed [18], [19].

B. Latency Components

The one-way URLLC latency (Υ) is defined from the time that a URLLC payload arrives at the network, until it is successfully decoded at the UE. If the UE correctly receives the packet in the first transmission, the latency equals the first transmission delay as:

$$\Upsilon = d_{fa,q} + d_{bsp} + d_{tx} + d_{uep}, \tag{1}$$

where $d_{fa,q}$ denotes the frame alignment and queuing delay. The payload transmission time is denoted by d_{tx}. Processing time at the network and the UE are represented by d_{bsp} and d_{uep}, respectively. The frame alignment delay is a uniformly distributed random variable taking values between zero and one TTI. The queuing delay accounts for the time where the the payload arrives at the base station until is considered for scheduling in the next upcoming TTI. The transmission time is a discrete random variable. Depending on the packet size, channel quality and scheduling strategy, d_{tx} varies from one to multiple TTIs. The processing times at the network (d_{bsp}) and the UE (d_{uep}) are assumed to be constants, equal to 2.75 and 4.5 OFDM symbols, respectively [20]. In case of failure, the packet is subject to additional retransmission delay(s), d_{HARQ}^{RTT}, until either it is decoded successfully or the maximum number of retransmissions is reached. In line with [10], the minimum retransmission delay of $d_{HARQ}^{RTT} = 4$ TTIs is assumed.

C. Problem Formulation

The objective is to maximize the network capacity of serving both URLLC and eMBB services. The URLLC capacity is defined as the maximum served average URLLC traffic L^{llc}, while still ensuring the packets are successfully delivered with the reliability of P_{target} within the given latency budget of T_{target} , expressed as $P(\Upsilon \leq T_{target}) \geq P_{target}$. For eMBB, maximizing the well-known Proportional-Fair (PF) utility function is assumed [21]. Dropping notations t and c for the ease of presentation, for a cell with D_{tot} PRBs, the resource allocation problem is formulated as:

$$\begin{aligned}
&\max_{b_{u/m}^{j}} \quad \sum_{u\in\mathbf{U}} a_u R_u^{llc} + \sum_{m\in\mathbf{M}} \log \bar{R}_m^{mbb}, \\
&\text{Sub. to:} \sum_{u\in\mathbf{U}} b_u^j + \sum_{m\in\mathbf{M}} b_m^j \leqslant 1,\ \forall j \in \{1,\cdots,D_{tot}\}, \\
&\qquad \sum_{j=1}^{D_{tot}} b_{u/m}^j \geq \min(R_{u/m}^{llc/mbb},1)\cdot b_{u/m}^{\min},\ \forall u,m, \\
&\qquad R_u^{llc} \leqslant Q_u^{llc}\ \ \forall u, \\
&\qquad b_{u/m}^j \in \{0,1\}\ \ \forall u,m,j,
\end{aligned} \tag{2}$$

where the binary variable b_i^j ($i \in \{u, m\}$, $j \in \{1, \cdots, D_{tot}\}$) indicates if the j-th PRB is allocated to i-th UE. The achievable rate of the u-th URLLC and the average throughput of m-th eMBB UEs are denoted by R_u^{llc} and $\bar{R}_m^{mbb}$, respectively. The minimum control channel overhead of the i-th UE is denoted by $b_i^{\min}$. The variable a_u is the u-th URLLC user QoS indicator chosen to satisfy the low-latency constraint. A larger a_u value indicates it is higher priority UE. Buffered data of the u-th URLLC user is represented by Q_u^{llc}. The first constraint in (2) ensures that each PRB is assigned to maximum one UE (single-user transmission). The second constraint guarantees that each scheduled UE has been assigned the minimum required number of PRBs to include the scheduling grant. Finally, the third constraints takes into account that the URLLC users have rather small amounts of buffered data to be served per scheduling interval. Problem (2) is a non-linear integer optimization can be solved using brute-force algorithm with complexity $\mathcal{O}\left(D_{tot}^{|\mathbf{U}|+|\mathbf{M}|}\right)$. This is too high complexity for practical network implementations as the URLLC scheduling decision needs to be taken every TTI on a fast basis.

III. Proposed Packet Scheduling Solution

A low-complexity packet scheduling algorithm that is aware of traffic, latency, control channel, HARQ, and radio channel is proposed as schematically presented in Fig.1. In line with [10]–[13], to reduce the queuing delay and enhance the reliability, URLLC payloads are scheduled first. After scheduling URLLC, eMBB traffic is served on the remaining PRBs.

A. URLLC Scheduling

URLLC payloads are scheduled in the following order.

Pending HARQ Retransmission: First, we assign the highest priority to HARQ retransmissions by scheduling them immediately over the set of PRBs with the highest CQI values. Thus, additional queuing delay is avoided as the payloads are already subjected to retransmission delay(s) of d_{HARQ}^{RTT}. By scheduling HARQ retransmissions over the best set of PRBs, we aim at increasing the reliability and minimizing the probability of further retransmissions.

Buffered URLLC Packets: Buffered URLLC payloads are scheduled thereafter. A low complexity time/frequency domain scheduler is applied as follows. First, the time-domain (TD) scheduler selects a group of UEs that can be fully scheduled over the available PRBs. Buffered payloads that are closer to the latency target (i.e. have lower latency budget) are prioritized by the TD scheduler. The number of required PRBs for each payload is estimated from the reported wide-band CQI. The selected UEs are thereafter scheduled by the FD scheduler.

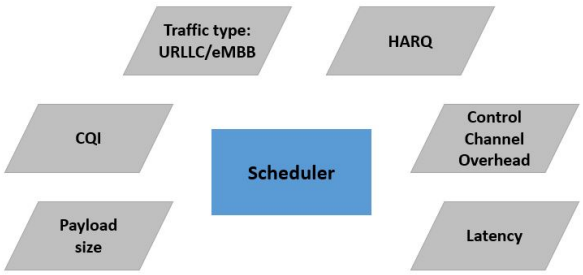

Fig. 1. Parameters affecting scheduling decision.

The FD scheduler utilizes multi-user radio channel-aware diversity mechanisms to achieve good performance. We utilize throughput to average (TTA) metric for scheduling URLLC payloads. Lets assume that r_u^p denotes the achievable throughput (TP) of PRB p for the u-th UE. The scheduler selects user $\hat{u}$ for being scheduled on PRB p which maximizes

$$\hat{u} = \max_u \frac{r_u^p}{\bar{\mathbf{r}}_u}, \tag{3}$$

where $\bar{\mathbf{r}}_u$ is the instantaneous full-bandwidth TP. Normalizing the achievable rate by the full-bandwidth TP, enhances fairness among the UEs and the probability to access to relatively good channels for all UEs [21]. As the rates of increase in TP is higher in low-SINR regimes [22], moderate and low-SINR UEs receive higher opportunity to occupy relatively better frequency-selective channel variations. Thus, scheduling based on (3) not only enhances the reliability of low-SINR UEs, but also fewer number of resources are needed to schedule the total payloads.

After UEs are scheduled in FD, the scheduler checks if it is possible to schedule more UEs on the remaining PRBs. The procedure is repeated until all buffered UEs are scheduled or there are not enough PRBs to schedule a full URLLC payload. For cases with insufficient PRBs for a full payload, at most one URLLC payload is segmented and scheduled over the remaining PRBs. To further reduce the cost of segmentation, UEs in good channel conditions (i.e. lower control channel overhead) are prioritized for segmentation. Details of the proposed scheduling is summarized in Algorithm 1.

B. eMBB Scheduling

After scheduling URLLC, eMBB UEs are scheduled on the remaining PRBs according to the PF metric. PRB p is assigned to UE $\hat{m}$ with the highest metric [21]

$$\hat{m} = \max_m \frac{r_m^p}{\bar{\mathbf{R}}_m}, \tag{4}$$

where $\bar{\mathbf{R}}_m$ is the m-th user average delivered throughput in the past, calculated by a moving average filter.

TABLE I
DEFAULT SIMULATION ASSUMPTIONS.

Description	Assumption
Environment	3GPP Urban Macro (UMa); 3-sector BSs with 500 meters inter-site distance. 21 cells.
Propagation	Urban Macro-3D.
Carrier	2 GHz (FDD), 20 MHz carrier bandwidth.
PHY numerology	30 kHz sub-carrier spacing configuration. PRB size of 12 sub-carrier (360 kHz).
TTI sizes	0.143 msec (4-symbols mini-slot).
MIMO	Single-user 2x2 closed loop single-stream (Rank-1) configuration. LMMSE-IRC receiver.
CSI	Periodic CSI every 5 msec, with 2 msec latency.
MCS	QPSK to 64 QAM, with same encoding rates as specified for LTE. Turbo codes.
Link adaptation	Dynamic MCS with 1% and 10% BLER for URLLC and eMBB, respectively.
HARQ	Asynchronous HARQ, Chase combining. HARQ-RTT=4 TTIs, max. 6 retransmissions.
User distribution	2100 URLLC and 210 eMBB UEs (Average 100 URLLC and 10 eMBB UEs per cell).
Traffic model	FTP3 downlink traffic with $B = 50$ bytes data for URLLC. Full buffer for eMBB.
Link-to-system (L2S) mapping	Based on MMIB mapping [23].

Algorithm 1 Proposed algorithm for URLLC packet scheduling

1: Schedule the HARQ retransmission over PRBs with the highest CQI values.
2: **while** Unscheduled UEs and enough PRBs **do**
3: Select a group of UEs with the lowest latency budget that can be fully scheduled.
4: For each selected UE and the available PRB, create pairs of UE/PRB and calculate the corresponding scheduling metric based on (3).
5: Sort pairs in the descending order of metric.
6: Allocate PRBs to UEs with the highest metric values, up to the required PRBs for each payload yields.
7: Remove if there is a segmented payload.
8: Update available PRBs.
9: **end while**
10: **if** Still unscheduled URLLC payload(s) and enough PRBs to partially schedule one payload **then**
11: Select the UE with the highest TP and schedule it over remaining PRBs.
12: **end if**

IV. SIMULATION RESULTS

A. Simulation Methodology

The performance of the proposed solution is evaluated by simulations using a highly detailed system level simulator that includes the 5G NR radio resource management functionalities as described in Section II. The simulation methodology is based on 3GPP 5G NR mathematical models and assumptions [1], [14], [24]. The assumed network configuration and default simulation parameters are summarized in Table I. At least five million URLLC packet transmissions are simulated to obtain statistical reliable results. This results in statistically reliable results with the confidence level of 95% for the 99.999% percentile of the latency [10]. For URLLC, the key performance indicator (KPI) is the one-way achievable latency with 99.999% reliability. For eMBB, the average cell TP is considered.

The results are compared against recent URLLC studies with PF scheduling [10], [11]. A comparison versus the well-known modified largest weighted delay first (M-LWDF) algorithm is also included. The M-LWDF scheduler is expressed as [21]

$$\hat{u} = \max_{u} \frac{-\log P_{target}}{T^{u}_{target}} d^{u}_{HOL} \frac{r^{p}_{u}}{\bar{\mathbf{R}}_{m}}, \tag{5}$$

where d^{u}_{HOL} is the head of line delay of user u. For both the PF and M-LWDF algorithms, URLLC UEs are scheduled first. eMBB traffic is served over the remaining PRBs. The network does not discard delayed packets.

B. Performance Results

Fig. 2 depicts the complementary cumulative distribution function (CCDF) of the URLLC latency for different offered URLLC loads from 4 to 14 Mbps/cell. At low offered loads, the latency performance is mainly affected by the transmission delay, processing times, and HARQ-RTT. URLLC payloads usually occupy only part of the available bandwidth and a few UEs compete for the resources. Thus, access to relatively good channels is possible for most UEs. Therefore, all scheduling methods have the same performance at low loads.

As the offered load increases, the queuing delay becomes more dominant. It is observed that the proposed solution provides significant latency improvement as the load increase. As an example, at 12 Mbps/cell load, the latency at 10^{-5} outage probability with PF, M-LWDF and the proposed algorithm is 4.5, 2.92 and 1.38 msec, respectively. This is equivalent to 70% and 53% latency gain in comparison with PF and M-LWDF scheduling, respectively. The proposed algorithm also shows a robust behaviour over the offered load variations, where the latency increases from 1.20 to 1.56 msec when the load is increased from 4 to 14 Mbps. In comparison, the latency increase corresponding to the

TABLE II
NETWORK PERFORMANCE FOR DIFFERENT URLLC OFFERED LOADS

Scenario			Offered URLLC load [Mbps]					
			4	**8**	**10**	**12**	**14**	**15**
URLLC latency at the outage probability of 10^{-5} [msec]	PF		1.21	1.5	2.3	4.5	69	358
	M-LWDF		1.2	1.36	1.63	2.92	10.45	22.5
	Proposed		1.2	1.24	1.31	1.38	1.56	2.27
	Relative improvement to	PF	0 %	18 %	57 %	70 %	98 %	99.4 %
		M-LWDF	0 %	9 %	20 %	53 %	84 %	90 %
Average eMBB cell throughput [Mbps/cell]	PF		34.6	25.3	20.8	16.2	11.5	9.2
	M-LWDF		34.6	25.3	20.8	16.3	11.64	9.3
	Proposed		34.7	25.6	21.3	17.07	12.9	10.8
	Relative improvement to	PF	0 %	1.3 %	2.5 %	5.3 %	12 %	17.3 %
		M-LWDF	0 %	1.3 %	2.5 %	4.7 %	11 %	16.1 %

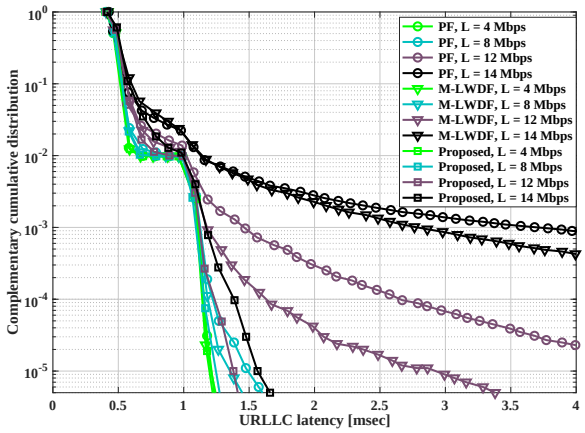

Fig. 2. URLLC latency distribution for different URLLC offered loads and scheduling methods.

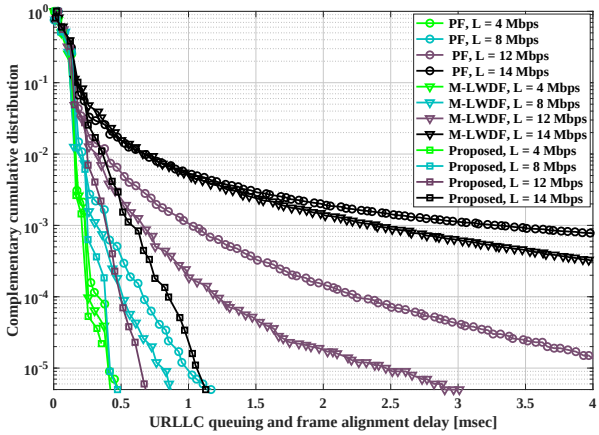

Fig. 3. Queuing and frame alignment delay for different offered loads and scheduling methods.

same load increase for the PF and M-LWDF algorithm is 1.21 to 69 msec and 1.20 to 10.45 msec, respectively.

Fig. 3 presents the CCDF of the combined queuing and frame alignment delay for different offered loads. As expected, the queuing delay increases with the offered load. The Figure shows the superior performance of the proposed algorithm in reducing the tail of the queuing delay which is important for URLLC traffic. For example, at 12 Mbps offered load only 0.01% of the payloads experience more than 0.5 msec queuing and frame alignment delay. While for M-LWDF and PF, it increases to 0.23% and 0.53%, respectively.

Table II presents the URLLC latency and the average eMBB cell TP for different scheduling and offered URLLC traffic settings. As the URLLC traffic is always prioritized over the eMBB, the average eMBB TP decreases when increasing the URLLC load. It can be seen from the table that the proposed solution improves both the URLLC latency and eMBB TP. At 14 Mbps URLLC load, it provides 98% URLLC latency reduction as well as 12% increase in eMBB TP in comparison to PF. Gains of 84% URLLC latency reduction and 11% eMBB TP enhancement are achieved over the M-LWDF. The performance benefits come as the results of: (i) considering the latency budget as the main scheduling parameters for URLLC (prioritizing UEs with the lowest latency budget). (ii) reducing the control channel overhead by single-TTI transmission of URLLC payloads, (iii) efficient FD multiplexing of URLLC UEs that results in fewer number of allocated resources to schedule the URLLC payloads.

V. CONCLUSION

We studied the problem of resource allocation for mixed URLLC and eMBB traffic in 5G NR multi-service networks. A latency-QoS, control channel, HARQ, and radio channel aware scheduling algorithm is proposed to enhance the performance of both URLLC and eMBB traffic. The proposed algorithm exploits the gains of frequency-selective multi-user scheduling while avoiding unnecessary and costly segmentation of URLLC payloads over multiple transmissions. The solution benefits from low computational complexity and is attractive for practical network implementation. Results show significant latency improvement of URLLC traffic as well as higher average eMBB throughput. As an example, at 14 Mbps URLLC offered load, the latency of URLLC at the 10^{-5} outage level is improved by 98% compared state of the art proportional fair scheduling and also the average eMBB throughput is increased by 12%.

ACKNOWLEDGEMENT

Part of this work has been performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union. The authors would like to acknowledge the contributions of their colleagues in the project, although the views expressed in this contribution are those of the authors and do not necessarily represent the project.

REFERENCES

[1] 3GPP Technical Specification 38.300, "NR and NG-RAN overall description; stage-2," Version 2.0.0, December 2017.

[2] 3GPP Technical Specification 23.501, "Technical specification group services and system aspects, system architecture for the 5G system," Release 15, December 2017.

[3] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk and scale," *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1801.01270.

[4] P. Popovski *et al.*, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, March 2018.

[5] K. I. Pedersen *et al.*, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53–59, March 2016.

[6] A. Destounis, G. S. Paschos, J. Arnau, and M. Kountouris, "Scheduling URLLC users with reliable latency guarantees," in *2018 16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2018, pp. 1–8.

[7] E. Khorov, A. Krasilov, and A. Malyshev, "Radio resource and traffic management for ultra-reliable low-latency communications," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[8] ——, "Radio resource scheduling for low-latency communications in LTE and beyond," in *2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS)*, June 2017, pp. 1–6.

[9] C.-P. Li *et al.*, "5G ultra-reliable and low-latency systems design," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.

[10] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28 912–28 922, 2018.

[11] ——, "Multiplexing of latency-critical communication and mobile broadband on a shared channel," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[12] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38 451–38 463, 2018.

[13] A. Anand, G. D. Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, April 2018, pp. 1970–1978.

[14] 3GPP Technical Report 38.913, "Study on scenarios and requirements for next generation access technologies," Version 14.1.0, March 2016.

[15] 3GPP Technical Report 38.802, "Study on new radio access technology physical layer aspects," Version 14.0.0, March 2017.

[16] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 210–217, March 2018.

[17] A. Karimi *et al.*, "Centralized joint cell selection and scheduling for improved URLLC performance," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, September 2018, pp. 1–6.

[18] K. I. Pedersen, S. R. Khosravirad, G. Berardinelli, and F. Frederiksen, "Rethink hybrid automatic repeat request design for 5G: Five configurable enhancements," *IEEE Wireless Communications*, vol. 24, no. 6, pp. 154–160, December 2017.

[19] D. Chase, "Code combining - a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Transactions on Communications*, vol. 33, no. 5, pp. 385–393, May 1985.

[20] 3GPP Technical Documents R1-1808449, "IMT-2020 self-evaluation: UP latency analysis for FDD and dynamic TDD with UE processing capability 2 (URLLC)," August 2018.

[21] F. Capozzi *et al.*, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 2, pp. 678–700, Second 2013.

[22] G. Ku and J. M. Walsh, "Resource allocation and link adaptation in LTE and LTE-Advanced: A tutorial," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1605–1633, third-quarter 2015.

[23] T. L. Jensen, S. Kant, J. Wehinger, and B. H. Fleury, "Fast link adaptation for MIMO OFDM," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 8, pp. 3766–3778, October 2010.

[24] IMT Vision, "Framework and overall objectives of the future development of IMT for 2020 and beyond," International Telecommunication Union (ITU), Document, Radiocommunication Study Groups, February 2015.